The ICNALE Global Rating Archives: A New Assessment Dataset for Learner Corpus Studies

Ishikawa, Shin'Ichiro

(Citation)
Learner Corpus Studies in Asia and the World,6:13-38

(Issue Date)
2024-03-20

(Resource Type)
departmental bulletin paper

(Version)
Version of Record

(JaLCDOI)
https://doi.org/10.24546/0100487713

(URL)
https://hdl.handle.net/20.500.14094/0100487713

The ICNALE Global Rating Archives
—A New Assessment Dataset for Learner Corpus Studies—

ISHIKAWA, Shin'ichiro (Kobe University)

Abstract

The scope of the learner corpus research can be further expanded by combining learner output data with the assessment data. Thus, the author developed the ICNALE Global Rating Archives (ICNALE GRA). The ICNALE GRA includes the rubric-based assessment data collected from 160 raters with varied L1, regional, and occupational backgrounds, who rated 140 speeches and the same number of essays produced by college students in Asia as well as L1 English native speakers. This paper first introduces the outline of the project and describes the content of the dataset. Finally, it touches upon the possibilities of utilising the GRA data for pedagogical and research purposes.

Keywords

ICNALE, Learner Corpus Studies, Assessment Data

1. Introduction

A greater emphasis has been put on the development of productive skills in recent English education in Asia, where teaching receptive skills was traditionally prioritised. Such a shift in the direction of English language teaching has made many teachers and learners in Asia face old and new questions on what a good speech/essay is and how it can be different from other speeches/essays.

With the aim of offering a reliable dataset to reconsider this question from an evidence-based perspective, the author recently released the Global Rating Archives v2 (GRA2) as a part of the International Corpus Network of Asian Learners of English (ICNALE) (Ishikawa, 2023a). The ICNALE GRA2 includes the ratings that eighty raters (160 in total) gave to 140 speeches and the same number of essays, both of which were produced by college students in ten countries and regions in Asia.

The GRA project began in April 2020. The project team first selected the rating samples and then developed the rating rubric, the rating sheet, and the rater registration system. In August 2021, the ICNALE GRA v.01 was released, which included the rating data from 80 raters (40 speech raters and 40 essay raters). In November 2022, the ICNALE GRA v1.0 was released, which included the rating data from 120 raters. Then, in October 2023, the ICNALE GRA v2.0 was released, which included the rating data from 160 raters (80 speech raters and 80 essay raters). After several minor corrections and updates, the ICNALE GRA v2.1 was released in March 2024.

2. ICNALE GRA Project

The aim and the principle of the ICNALE GRA project are explained in Ishikawa (2020), and the detail of the GRA data collection scheme is described in the ICNALE official guidebook (Ishikawa, 2023a, pp. 61–70). Based on these references, this section briefly outlines the project.

2.1 Rating Samples

Rating samples were the 140 speeches and the same number of essays produced by college students in ten countries and regions in Asia as well as L1 English native speakers. The speech data was taken from the ICNALE Spoken Dialogues, and the essay data was from the ICNALE Written Essays (See Sections 3.2 and 3.4 of the ICNALE guidebook for these two data modules). The details of the rating samples are shown in Table 1.

Table 1

Details of the Speeches and Essays Rated in the GRA Project

	Speeches	Essays
Task	Initial 90 seconds of the roleplays where participants were requested to persuade their college supervisors to permit them to continue doing their current part-time jobs.	200-300-word essays written about the topic: "It is important for college students to have a part-time job."
Data offered to raters	Sound files (available from the online server)	Text files
Speakers/ Writers	College students in EFL regions: 120 20 from China (CHN), Indonesia (IDN), Japan (JPN), Korea (KOR), Taiwan (TWN), and Thailand (THA) College students in ESL regions: 16 Four from Hong Kong (HKG), Pakistan (PAK), The Philippines (PHL), and Singapore (SIN) [essays] or Malaysia (MYS) [speeches] L1 English native speakers (ENS): Four Total: 140	

The principal target of the project was learners of English as a foreign language (EFL) in Asia who study English as a foreign language at schools and have relatively limited opportunities for L2 use in their daily lives. For comparison, however, a small number of the outputs of learners of English as a second language (ESL) in Asia and L1 English native speakers (ENS) were also added.

All samples were randomised and anonymised, meaning that raters needed to assess each sample without knowing the nationality and English proficiency level of each speaker and writer.

2.2 Rating Rubric

The raters were asked first to assess the quality of a sample from a holistic viewpoint (100 points) and then to carefully evaluate each of the ten rating criteria (10 points each), which were subdivided into three rating categories: language, content, and attitude. Finally, the overall rating score (ORS) was calculated by averaging the holistic scores (100 points) and the analytic score sum (ANAS) (100 points).

They were also asked to rate each sample from the viewpoint of English as a lingua franca (ELF), a type of English used for communication mainly between non-native speakers rather than the so-called native-speaker norm. Details of a rubric and a scoring standard are explained in a rater guidebook (See the Appendix).

Table 2

Rating Viewpoint of Each Criterion

Criterion	Rating Viewpoint
Holistic	To what extent do you think the sample is close to an ideal ELF speech/essay?
Intelligibility	To what extent can you "decode," namely, verbally understand what is said/written? Intelligibility, which concerns the understandability of the language, should be discriminated from comprehensibility, which concerns the understandability of the content.
Complexity	To what extent do you think the speaker/ writer uses morphologically and/or semantically complex words, phrases, expressions, constructions, and grammar?
Accuracy	To what extent do you think the sample is error-free in terms of vocabulary and grammar?

Fluency	To what extent do you think the speaker/writer is fluent in the speeches/ essays?
Comprehensibility	To what extent can you understand the content of the speech/essay?
Logicality	To what extent do you think the idea presented in the speech/essay is logical and reasonable?
Sophistication	To what extent do you think the ideas presented in the speech/essay are well-sophisticated, critically thought, unique, original, and innovative?
Purposefulness	To what extent do you think the speaker/writer consistently and consciously pays attention to the purpose of the task?
Willingness	To what extent do you think the speaker/writer is willing to communicate?
Involvement	To what extent do you think the participant tries to make the hearer/reader involved in his/her discourse rather than speaking/writing one-sidedly?

Raters were told to enter the scores on the Excel rating sheet prepared for this project and to confirm that the averages and the standard deviations of their scores, which were automatically calculated and displayed on the sheet, should fall between 4-6 per 10 points or 45 and 55 per 100 points, and between 2-3 per 10 points or 20-30 per 100 points, respectively. This is a measure to exclude assessments that are too lenient, too strict, or too flat (i.e., giving the same score to all the samples).

2.3 Rater Training

All the raters were told to carefully read the rater guide (See Appendix) and understand the project aim and the rating policy, and then they were told to take an online check test. Only after passing the test were they permitted to commence their rating task.

2.4 Rater Backgrounds

Considering that any output may be heard or read by a wide variety of people worldwide, it would not be appropriate to collect assessment data only from some authority. Thus, the project team decided to prioritise the "collective intelligence" of a wider variety of people rather than the judgment of a single expert or a few experts.

The project team finally recruited 80 speech raters and the same number of essay raters (160 raters in total), all of whom are ELF users using English for their professional purposes. Among the 160 raters, 59 assessed both speeches and essays, and the remaining 42 assessed either of them.

Table 3

Rater Backgrounds

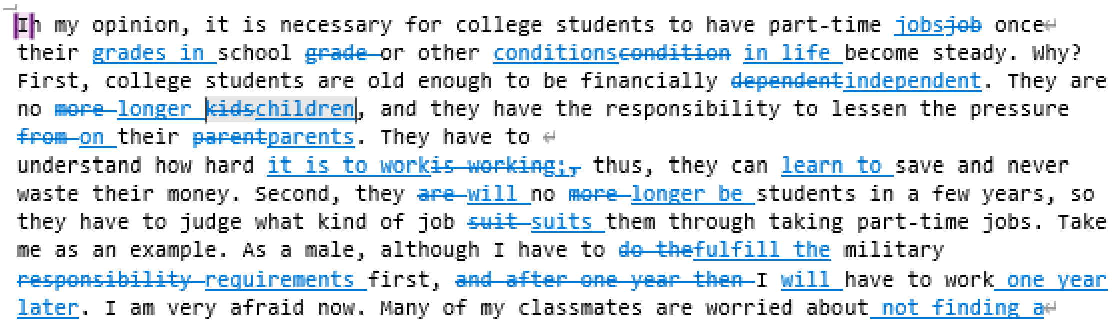
	Speech Raters (80 persons)	Essay Raters (80 persons)
Age	20s (20), 30s (36), 40s (17), 50s (3), 60s (3), 70s (1)	20s (16), 30s (40), 40s (15), 50s (5), 60s (3), 70s (1)
Sex	F (40), M (30)	F (49), M (31)
Degree	HS (1), BA (29), MA (35), Dr. (15)	HS (2), BA (25), MA (38), Dr. (15)
Proficiency (CEFR)*	B1 (2), B2 (19), C1 (33), C2 (14), (Near)Native (12)	B1 (4), B2 (17), C1 (32), C2 (12), (Near)Native (15)
Occupation	Business (10), NA (8), Others (10), English Teachers (39), Other Subject Teachers (12), Translators (1)	Business (11), NA (4), Others (8), English Teachers (44), Other Subject Teachers (11), Translators (1)
Expert Fields	English (31), Languages (Not English) (3), Humanities (7), Life Sciences (3), Natural Sciences (10), Social Sciences (26)	English (36), Languages (Not English) (5), Humanities (7), Life Sciences (2), Natural Sciences (8), Social Sciences (22)
Rating Experiences	Never (12), 1-5 times (21), 6+ times (47)	Never (10), 1-5 times (17), 6+ times (53)
L1	Arabic (1), Chinese (10), English (6), Filipino/Tagalog (13), French (1), German (1), Hindi (1), Hmong (1), Indonesian (6), Japanese (9), Konkani (1), Korean (3), Lao (10), Malay (2), Portuguese (1), Punjabi (2). Sinhala (1), Thai (6), Urdu (2), Uyghur (1), Vietnamese (1), Yoruba (1)	Arabic (1), Bangla (1), Chinese (5), English (10), Filipino/Tagalog (11), French (1), German (1), Hmong (1), Indonesian (5), Japanese (12), Konkani (1), Korean (3), Lao (12), Malay (3), Punjabi (1). Sinhala (1), Spanish (1), Thai (6), Urdu (3), Vietnamese (1)

2.5 Editing Data

In order to offer additional data to discuss the linguistic quality of learners' written outputs, the project team hired professional proofreaders and asked them to edit all 140 essays. As in the data collected in the ICNALE Edited Essays (See Section 3.5 of the ICNALE guidebook), editing was done on MS Word, which means that a user can easily see how many words and what words were inserted or deleted.

Figure 1

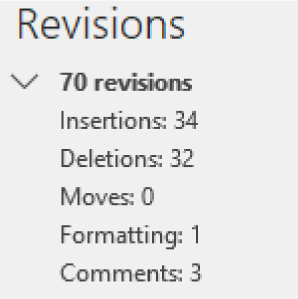
A Part of the Edited Essay



In my opinion, it is necessary for college students to have part-time jobs~~job~~ once
their grades in school ~~grade~~ or other conditions~~condition~~ in life become steady. Why?
First, college students are old enough to be financially ~~dependent~~independent. They are
no ~~more~~ longer ~~kids~~children, and they have the responsibility to lessen the pressure
~~from~~ on their ~~parent~~parents. They have to
understand how hard it is to work~~is working~~;~~,~~ thus, they can learn to save and never
waste their money. Second, they ~~are~~ will no ~~more~~ longer be students in a few years, so
they have to judge what kind of job ~~suit~~ suits them through taking part-time jobs. Take
me as an example. As a male, although I have to ~~do the~~fulfill the military
~~responsibility~~ requirements first, ~~and after one year then~~ I will have to work one year
later. I am very afraid now. Many of my classmates are worried about not finding a

Figure 2

Revision Summary Seen in the MS Word Reviewing Pane



Revisions

70 revisions

Insertions: 34
Deletions: 32
Moves: 0
Formatting: 1
Comments: 3

3. ICNALE GRA Dataset

3.1 Distribution

The ICNALE GRA is available from the ICNALE website (language.sakura.ne.jp/icnale/). After registration, users can download the Excel datasheet to their computers.

3.2 Contents of the Dataset

The dataset (v2.1) includes the Excel data file, rating samples, and the related document. The "Rating Samples" folder contains (1) 140 essay samples used for rating,

(2) the same number of edited versions by professional proofreaders, and (3) 140 speech samples (only the links to the online sound files).

3.3 Datafile

This section explains the type of data offered in the Excel data file, which consists of six sheets.

Figure 3

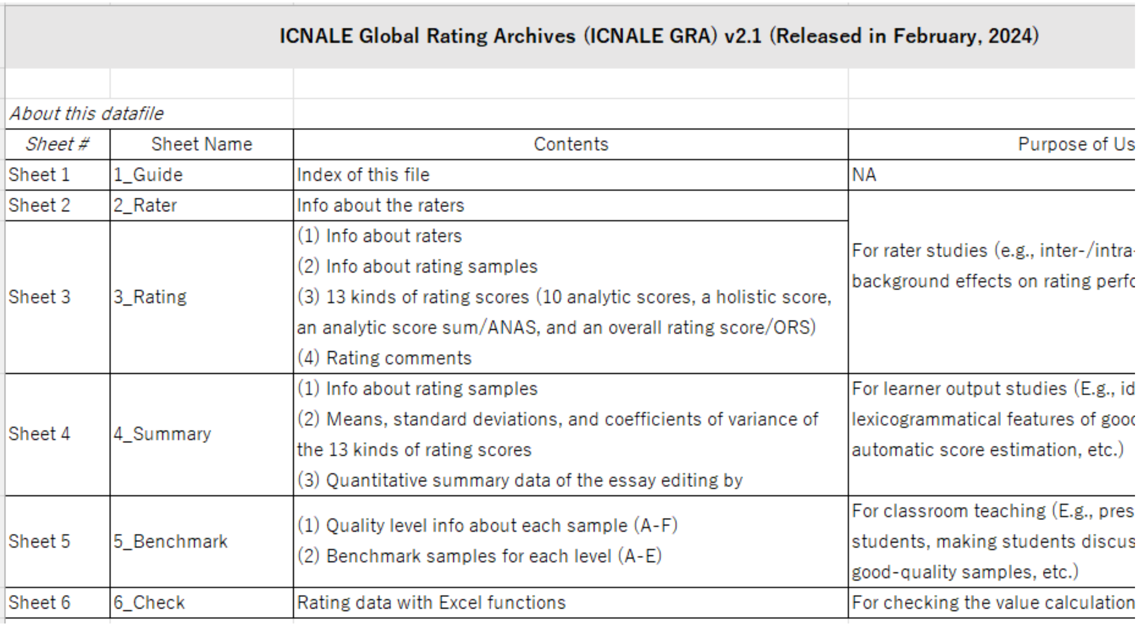
Six Sheets included in the GRA Datafile

3.3.1 Guide Sheet

The guide sheet introduces the contents of each sheet.

Figure 4

The Guide Sheet Showing the Contents and Purpose of Use of Each Sheet



ICNALE Global Rating Archives (ICNALE GRA) v2.1 (Released in February, 2024)

About this datafile

Sheet #	Sheet Name	Contents	Purpose of Us
Sheet 1	1_Guide	Index of this file	NA
Sheet 2	2_Rater	Info about the raters	For rater studies (e.g., inter-/intra- background effects on rating perfc
Sheet 3	3_Rating	(1) Info about raters (2) Info about rating samples (3) 13 kinds of rating scores (10 analytic scores, a holistic score, an analytic score sum/ANAS, and an overall rating score/ORS) (4) Rating comments	
Sheet 4	4_Summary	(1) Info about rating samples (2) Means, standard deviations, and coefficients of variance of the 13 kinds of rating scores (3) Quantitative summary data of the essay editing by	For learner output studies (E.g., id lexicogrammatical features of gooc automatic score estimation, etc.)
Sheet 5	5_Benchmark	(1) Quality level info about each sample (A-F) (2) Benchmark samples for each level (A-E)	For classroom teaching (E.g., pres students, making students discus good-quality samples, etc.)
Sheet 6	6_Check	Rating data with Excel functions	For checking the value calculation

This sheet also offers information about the principal investigator, funding, and citation guide.

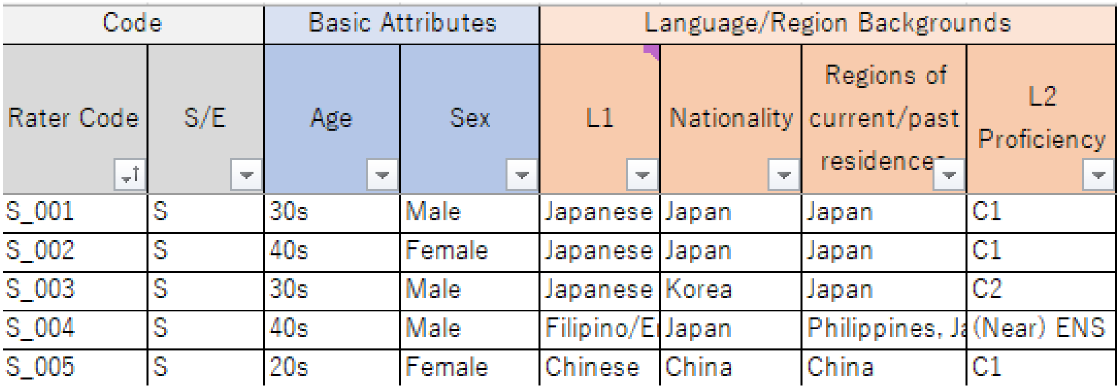
3.3.2 Rater Sheet

The rater sheet introduces detailed background information about each of the 160 raters who joined the project. The information is presented in five categories.

Figure 5 shows the first three categories.

Figure 5

Information Shown in the Categories of "Code," "Basic Attributes," and "Language/Region Backgrounds"



Code		Basic Attributes		Language/Region Backgrounds			
Rater Code	S/E	Age	Sex	L1	Nationality	Regions of current/past residence	L2 Proficiency
S_001	S	30s	Male	Japanese	Japan	Japan	C1
S_002	S	40s	Female	Japanese	Japan	Japan	C1
S_003	S	30s	Male	Japanese	Korea	Japan	C2
S_004	S	40s	Male	Filipino/E	Japan	Philippines, J	(Near) ENS
S_005	S	20s	Female	Chinese	China	China	C1

"Code" shows the rater code (e.g., S_001, E_001) and the type of the outputs that each rater assessed (S: Speeches or E: Essays). Regarding the rater code, the same number does not mean the same person.

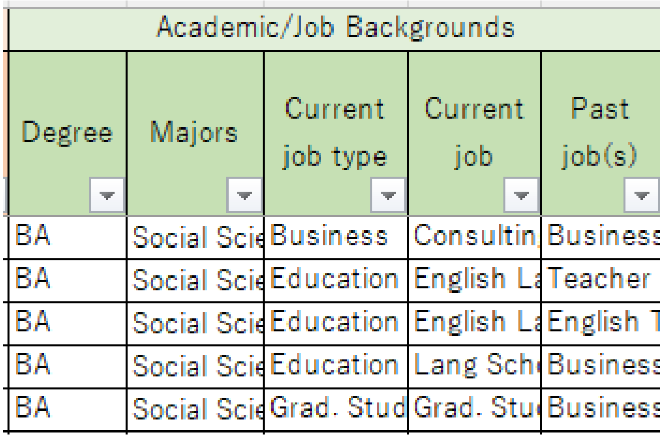
"Basic Attributes" introduces the age (e.g., 20s, 60s) and the sex (Female or Male) of each rater.

"Language/Region Backgrounds" shows (1) the rater's L1, (2) nationality, (3) regions of current/past residences (e.g., China, China and Japan), and (4) L2 proficiency levels based on the CEFR. Regarding (4), each rater was requested to choose one from B1, B2, C1, C2, and (near) native after reading the can-do statement of each level.

Next, Figure 6 shows the category of "Academic/Job Backgrounds."

Figure 6

Information Shown in the Category of "Academic/Job Backgrounds"



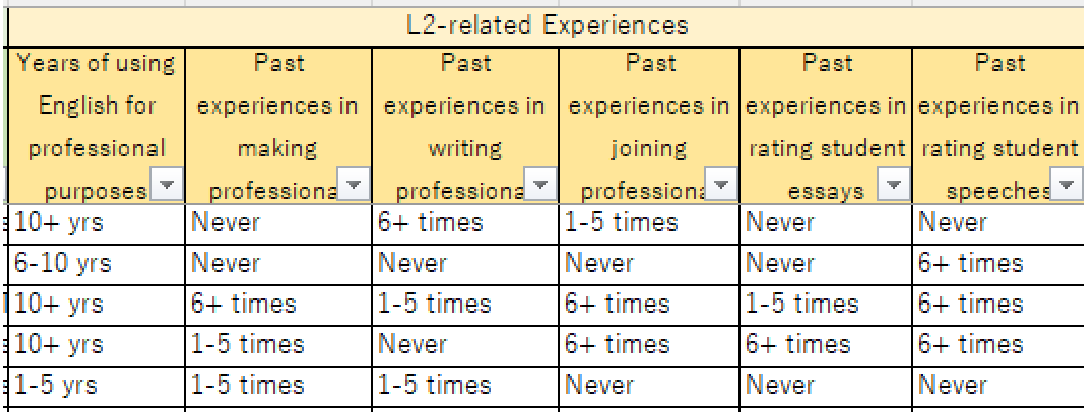
Academic/Job Backgrounds				
Degree	Majors	Current job type	Current job	Past job(s)
BA	Social Scie	Business	Consultin	Business
BA	Social Scie	Education	English La	Teacher
BA	Social Scie	Education	English La	English T
BA	Social Scie	Education	Lang Sch	Business
BA	Social Scie	Grad. Stud	Grad. Stu	Business

This category offers information about each rater's (1) highest degree (e.g., BA, MA), (2) academic major at a college, (3) current job type, (4) detailed description of the current job, and (5) past job history. Regarding (2), a rater was requested to choose one from the six fields: English, Languages (not English), Humanities, Social Sciences, Natural Sciences, and Life Sciences. Then, regarding (3), they chose one from the five fields: Education (English), Education (not English), Business, Graduate Students, and Others.

Then, Figure 7 shows the category of "L2-related Experiences."

Figure 7

Information Shown in the Category of "L2-related Experiences"



L2-related Experiences					
Years of using English for professional purposes	Past experiences in making professiona	Past experiences in writing professiona	Past experiences in joining professiona	Past experiences in rating student essays	Past experiences in rating student speeches
10+ yrs	Never	6+ times	1-5 times	Never	Never
6-10 yrs	Never	Never	Never	Never	6+ times
10+ yrs	6+ times	1-5 times	6+ times	1-5 times	6+ times
10+ yrs	1-5 times	Never	6+ times	6+ times	6+ times
1-5 yrs	1-5 times	1-5 times	Never	Never	Never

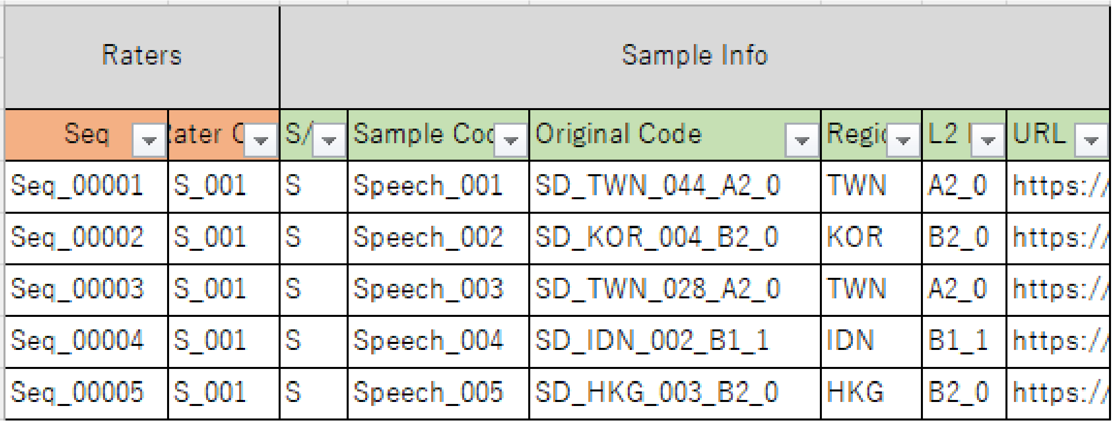
This category shows (1) how long a rater has used English for their professional purposes (e.g., 1-5 years, 10+ years), (2-4) how often they have made presentations, written documents and articles, and joined discussions in English as a part of their business (e.g., Never, 1-5 times), and (5-6) how often they have rated student essays and speeches (e.g., Never, 6+times).

3.3.3 Rating Sheet

The rating sheet includes information about raters, rating samples, rating scores, and rating comments. Figure 8 shows the first two categories.

Figure 8

Information Shown in the Category of "Rater" and "Sample Info"



Raters		Sample Info						
Seq	ater C	S/	Sample Coc	Original Code	Regi(	L2 (	URL
Seq_00001	S_001	S	Speech_001	SD_TWN_044_A2_0	TWN	A2_0	https://
Seq_00002	S_001	S	Speech_002	SD_KOR_004_B2_0	KOR	B2_0	https://
Seq_00003	S_001	S	Speech_003	SD_TWN_028_A2_0	TWN	A2_0	https://
Seq_00004	S_001	S	Speech_004	SD_IDN_002_B1_1	IDN	B1_1	https://
Seq_00005	S_001	S	Speech_005	SD_HKG_003_B2_0	HKG	B2_0	https://

"Raters" shows the sequential code of each rating data (Seq_00001 to Seq_22400) and the rater code. As each of the 160 raters assesses the set of 140 samples, the total number of ratings amounts to 22,400.

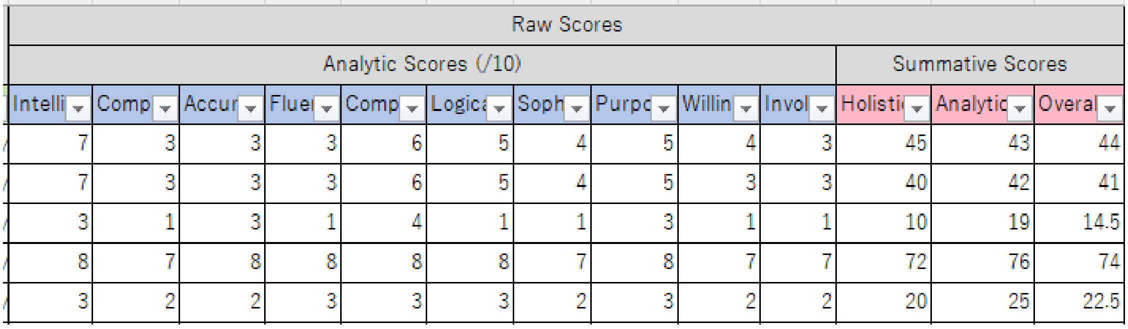
"Sample Info" introduces (1) the type of rating samples (S or E), (2) the code given to each rating sample (e.g. Speech_001, Essay_001), (3) the original sample code adopted in the ICNALE modules (e.g., SD_TWN_044_A2_0, WE_TWN_005_B1_2), (4) the participant's region of residence (e.g., TWN, CHN), (5) their L2 proficiency level based on the CEFR (A2, B1_1, B1_2, and B2+), which was estimated from their scores in the standard English proficiency tests such as TOEIC and TOEFL or the standard vocabulary size test (See Ishikawa 2023a, pp. 25–28), and (6) the links to the speech data stored online.

Regarding (2), the first two letters represent the ICNALE module type (SD: Spoken Dialogues or WE: Written Essays), the following three letters represent the participant's region (e.g., CHN, PHL), the next three digits show the participant code, and the last four letters represent the participant's CEFR levels.

Next, Figure 9 shows the categories of "Analytic Scores" and "Summative Scores," both of which are the unadjusted raw scores.

Figure 9

Information Shown in the Categories of "Analytic Scores" and "Summative Scores"



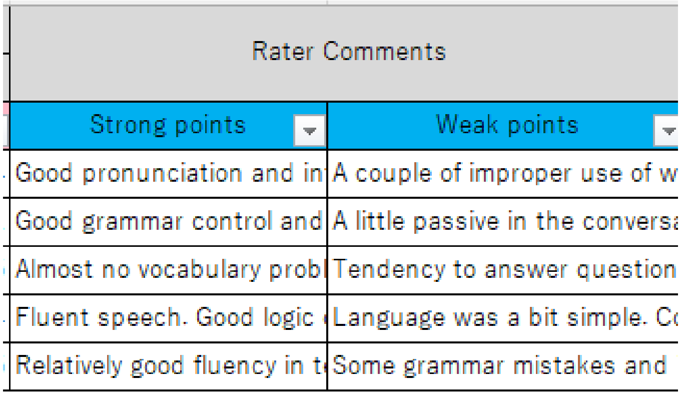
Raw Scores												
Analytic Scores (/10)										Summative Scores		
Intelli	Comp	Accur	Fluer	Comp	Logica	Soph	Purpo	Willin	Invol	Holisti	Analytic	Overal
7	3	3	3	6	5	4	5	4	3	45	43	44
7	3	3	3	6	5	4	5	3	3	40	42	41
3	1	3	1	4	1	1	3	1	1	10	19	14.5
8	7	8	8	8	8	7	8	7	7	72	76	74
3	2	2	3	3	3	2	3	2	2	20	25	22.5

"Analytic Scores" shows ten kinds of analytic scores (/10) assigned by each rater, and "Summative Scores" shows a holistic score (/100), an analytic score sum (ANAS) (/100), which was calculated by summing up the ten analytic scores, and an overall rating score (ORS) (/100), which was the mean of the holistic score and the ANAS.

Finally, Figure 10 shows the category of "Rater Comments."

Figure 10

Information Shown in the Category of "Rater Comments"



Rater Comments	
Strong points	Weak points
Good pronunciation and in	A couple of improper use of w
Good grammar control and	A little passive in the convers
Almost no vocabulary prob	Tendency to answer question
Fluent speech. Good logic	Language was a bit simple. C
Relatively good fluency in t	Some grammar mistakes and

"Rater Comments" introduces the rater's short remarks about the strong and weak points of each sample.

3.3.4 Summary Sheet

The summary sheet includes information about rating samples and three types of scores: means, standard deviations of the means (SD), and coefficients of variance of the means (CV).

Although SD is a widely used index showing the amount of variation in scores, it is easily influenced by the data units (e.g., 10-point scale or 100-point scale). Meanwhile, CV, which is obtained by dividing SD with means, is not influenced by the data units.

Table 4 presents the sample data for explaining the relationship between mean, SD, and CV, where A-E represent raters, and x-z represent the rating samples. The mean values are the same (7/10 or 70/100) for all three rating samples, and the holistic scores are all just ten times as large as analytic scores (7 to 70, 6 to 60, 2 to 20, etc.).

Table 4

Sample Rating Data to Show the Relationship between Mean, SD, and CV

Score		A	B	C	D	E	Mean	SD	CV
Analytic (/10)	x	7	7	7	7	7	7	0.00	0.00
	y	8	6	7	7	7	7	0.63	0.09
	z	10	4	10	2	9	7	3.35	0.48
Holistic (/100)	x	70	70	70	70	70	70	0.00	0.00
	y	80	60	70	70	70	70	6.32	0.09
	z	100	40	100	20	90	70	33.47	0.48

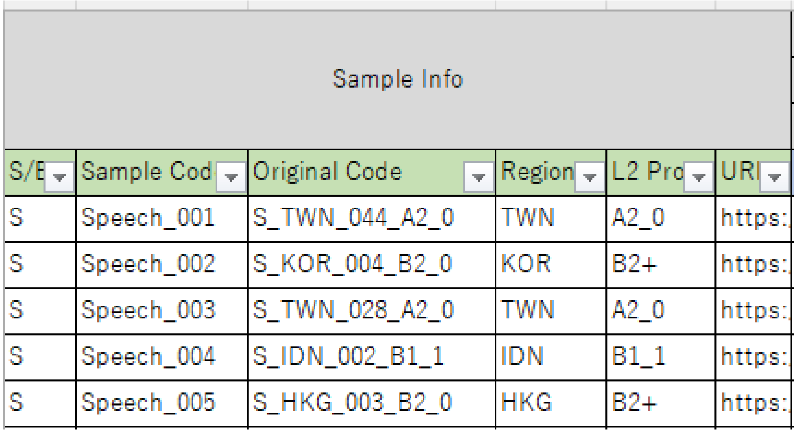
We compare the analytic scores assigned to the three samples. First, five raters assigned the same score (7) to x; therefore, the SD is calculated as zero (zero variation), and naturally, CV also becomes zero (0/7). Second, raters assigned relatively similar scores (6-8) to the sample y, and in this case, the SD is calculated as 0.63, and the CV becomes 0.09 (0.63/7). Last, raters assigned quite different scores (2-10) to the sample z, and in this case, the SD is calculated as 3.35, and the CV becomes 0.48 (3.35/7). As seen here, both SD and CV represent how (un)stably each sample has been rated.

Next, we pay attention to the holistic scores. The SD of the sample y, for example, now becomes 6.32, which is just ten times as large as 0.63. This exemplifies that the SD is relative to the units or the denominators. Meanwhile, the CV values remain the same (0.09). Thus, CV can be interpreted as a stable index measuring the absolute degree of variation in scores.

Figure 11 shows the category of “Sample Info.”

Figure 11

Information in the Category of "Sample Info"



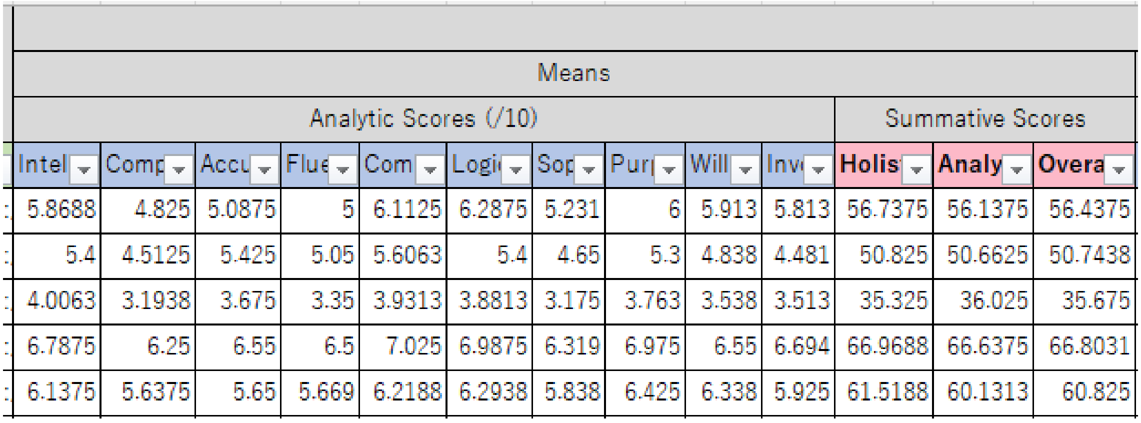
Sample Info					
S/E	Sample Cod	Original Code	Region	L2 Pro	URI
S	Speech_001	S_TWN_044_A2_0	TWN	A2_0	https:
S	Speech_002	S_KOR_004_B2_0	KOR	B2+	https:
S	Speech_003	S_TWN_028_A2_0	TWN	A2_0	https:
S	Speech_004	S_IDN_002_B1_1	IDN	B1_1	https:
S	Speech_005	S_HKG_003_B2_0	HKG	B2+	https:

"Sample Info" introduces information about (1) the type of rating samples (S/E), (2) the code given to each rating sample, (3) the original sample code adopted in the ICNALE modules, (4) the participant's region of residence, (5) their L2 proficiency level based on the CEFR, and (6) the links to the speech data stored online, all of which are the same with information presented in the rating sheet.

Then, Figures 12 –15 introduce the categories of "Means," "Standard Deviations," and "Coefficients of variance."

Figure 12

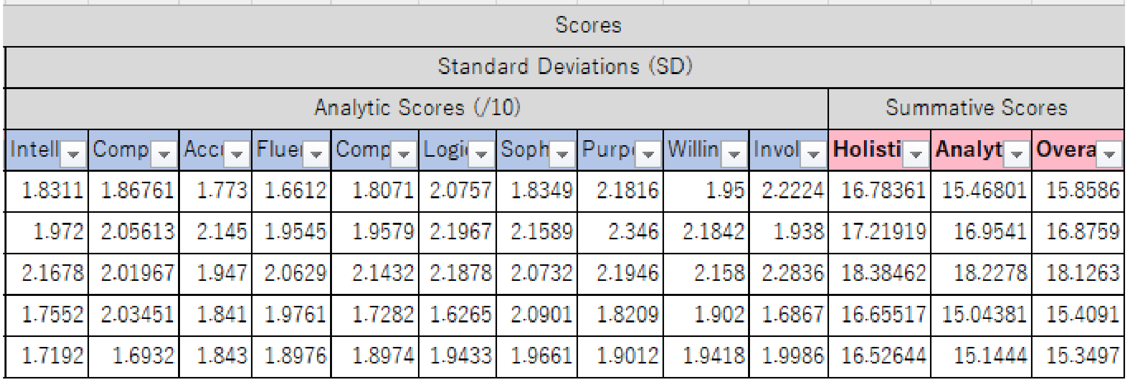
Information in the Category of "Means"



Means												
Analytic Scores (/10)										Summative Scores		
Intel	Comp	Accu	Flue	Com	Logi	Sop	Pur	Will	Inv	**Holis**	**Analy**	**Overa**
5.8688	4.825	5.0875	5	6.1125	6.2875	5.231	6	5.913	5.813	56.7375	56.1375	56.4375
5.4	4.5125	5.425	5.05	5.6063	5.4	4.65	5.3	4.838	4.481	50.825	50.6625	50.7438
4.0063	3.1938	3.675	3.35	3.9313	3.8813	3.175	3.763	3.538	3.513	35.325	36.025	35.675
6.7875	6.25	6.55	6.5	7.025	6.9875	6.319	6.975	6.55	6.694	66.9688	66.6375	66.8031
6.1375	5.6375	5.65	5.669	6.2188	6.2938	5.838	6.425	6.338	5.925	61.5188	60.1313	60.825

Figure 13

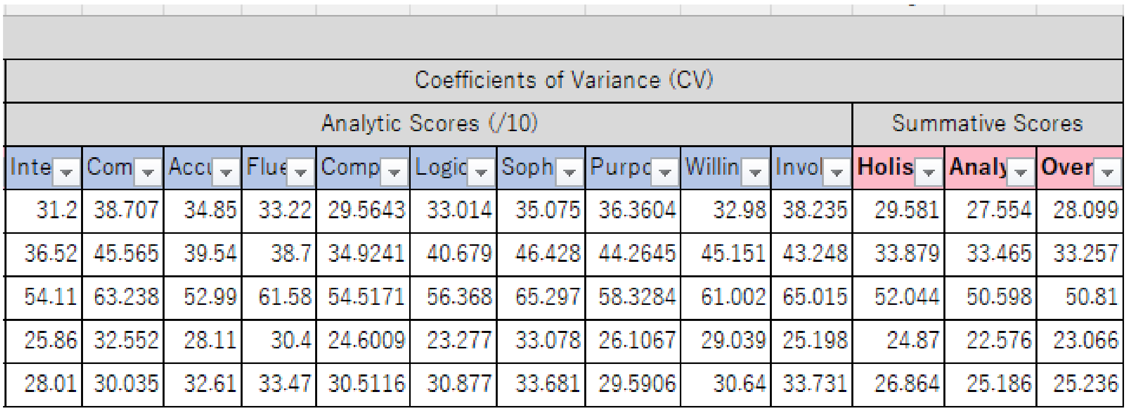
Information in the Category of "Standard Deviations"



Scores												
Standard Deviations (SD)												
Analytic Scores (/10)										Summative Scores		
Intell	Comp	Acc	Flue	Comp	Logi	Soph	Purp	Willin	Invol	**Holisti**	**Analyt**	**Overa**
1.8311	1.86761	1.773	1.6612	1.8071	2.0757	1.8349	2.1816	1.95	2.2224	16.78361	15.46801	15.8586
1.972	2.05613	2.145	1.9545	1.9579	2.1967	2.1589	2.346	2.1842	1.938	17.21919	16.9541	16.8759
2.1678	2.01967	1.947	2.0629	2.1432	2.1878	2.0732	2.1946	2.158	2.2836	18.38462	18.2278	18.1263
1.7552	2.03451	1.841	1.9761	1.7282	1.6265	2.0901	1.8209	1.902	1.6867	16.65517	15.04381	15.4091
1.7192	1.6932	1.843	1.8976	1.8974	1.9433	1.9661	1.9012	1.9418	1.9986	16.52644	15.1444	15.3497

Figure 14

Information in the Category of "Coefficients of Variance"



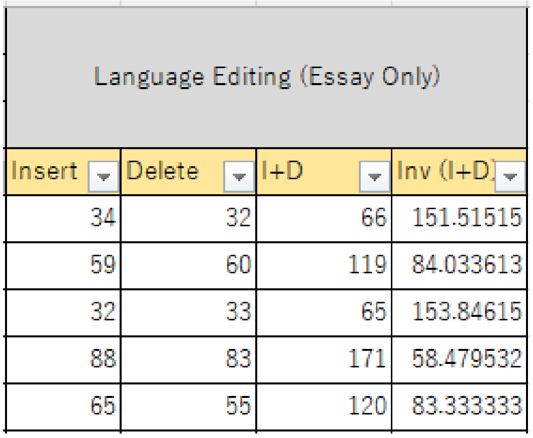
Coefficients of Variance (CV)												
Analytic Scores (/10)										Summative Scores		
Inte	Com	Acc	Flue	Comp	Logic	Soph	Purpo	Willin	Invol	**Holis**	**Analy**	**Over**
31.2	38.707	34.85	33.22	29.5643	33.014	35.075	36.3604	32.98	38.235	29.581	27.554	28.099
36.52	45.565	39.54	38.7	34.9241	40.679	46.428	44.2645	45.151	43.248	33.879	33.465	33.257
54.11	63.238	52.99	61.58	54.5171	56.368	65.297	58.3284	61.002	65.015	52.044	50.598	50.81
25.86	32.552	28.11	30.4	24.6009	23.277	33.078	26.1067	29.039	25.198	24.87	22.576	23.066
28.01	30.035	32.61	33.47	30.5116	30.877	33.681	29.5906	30.64	33.731	26.864	25.186	25.236

Each of the three categories shows the values of ten kinds of analytic scores (/10) and three types of summative scores: a holistic score (/100), an analytic score sum (ANAS) (/100), and an overall rating score (ORS) (/100).

Lastly, Figure 15 shows the category of "Language Editing," which was given only to essays.

Figure 15

Information in the Category of "Language Editing"



Language Editing (Essay Only)			
Insert	Delete	I+D	Inv (I+D)
34	32	66	151.51515
59	60	119	84.033613
32	33	65	153.84615
88	83	171	58.479532
65	55	120	83.333333

"Language Editing" introduces (1) the number of insertions, (2) the number of deletions, (3) the total number of insertions and deletions (I+D), and (4) the reciprocal number (i.e., the inverse number) of (3), which was presented as ten-thousand-fold (10k) values.

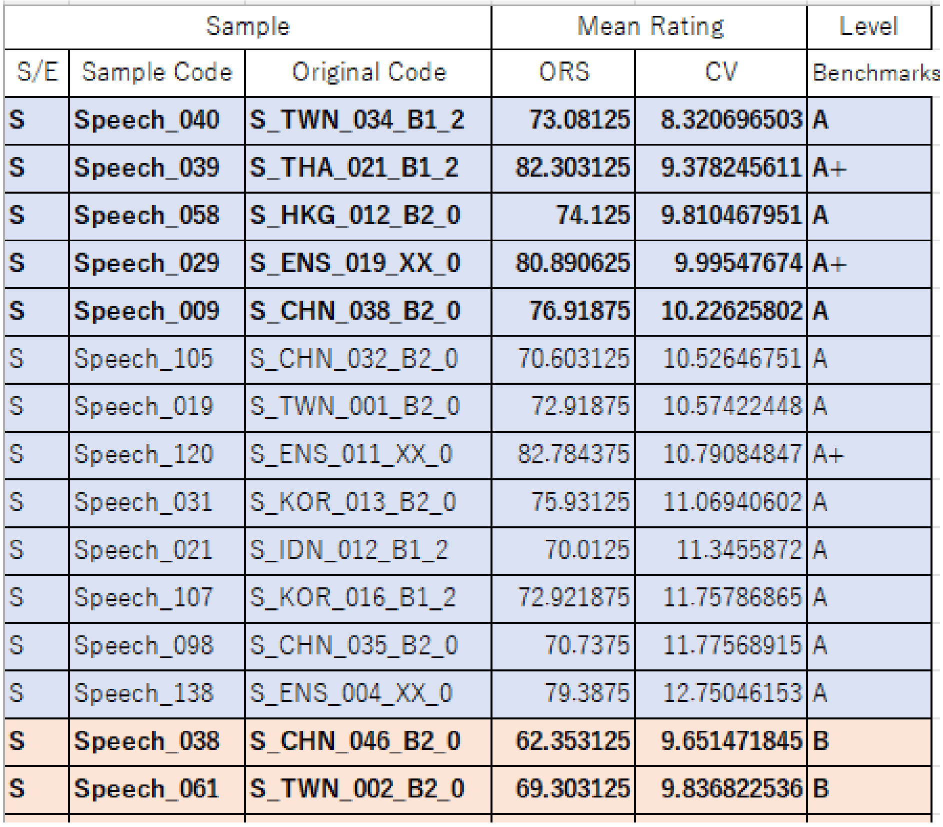
Contrary to the rating scores, the number of edits, which tends to be larger if a sample includes many lexical and grammatical problems, cannot be an index of the "goodness" of a sample. Meanwhile, the invert value can be regarded as an index of the goodness of a sample. In the case of Essay_001, the number of insertions and deletions are 34 and 32, respectively. Therefore, the number of edits (I+D) becomes 66 (34+32). Then, the inverse number is calculated as 0.015151 (1/66), which becomes 151.51 as a ten-thousand-fold value.

3.3.5 Benchmark Sheet

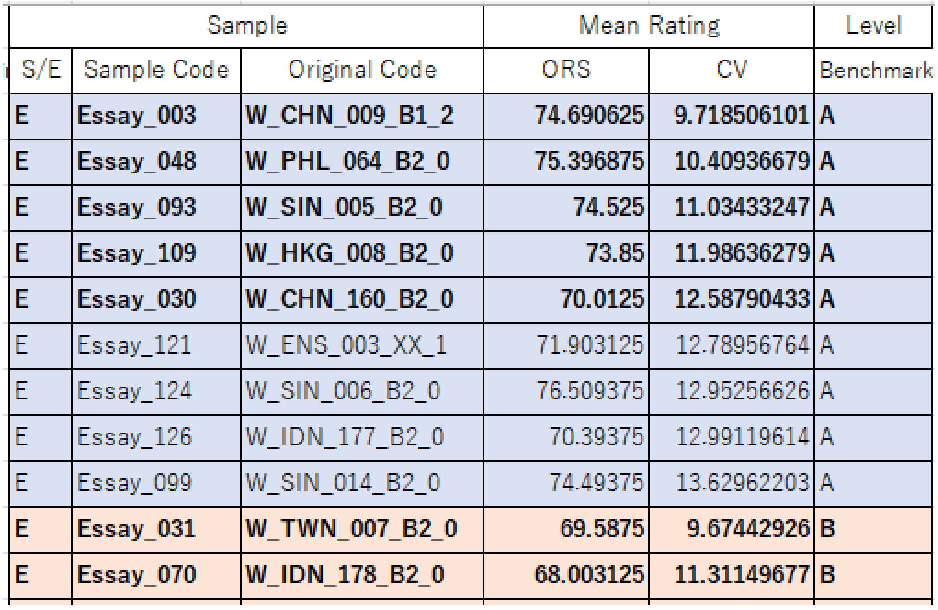
The benchmark sheet shows the quality level of each sample, and it introduces five benchmark samples for each level. Figure 16 shows spoken and written samples classified in terms of the quality level.

Figure 16

Information in the Benchmarks Sheet



Sample			Mean Rating		Level
S/E	Sample Code	Original Code	ORS	CV	Benchmarks
S	**Speech_040**	**S_TWN_034_B1_2**	**73.08125**	**8.320696503**	**A**
S	**Speech_039**	**S_THA_021_B1_2**	**82.303125**	**9.378245611**	**A+**
S	**Speech_058**	**S_HKG_012_B2_0**	**74.125**	**9.810467951**	**A**
S	**Speech_029**	**S_ENS_019_XX_0**	**80.890625**	**9.99547674**	**A+**
S	**Speech_009**	**S_CHN_038_B2_0**	**76.91875**	**10.22625802**	**A**
S	Speech_105	S_CHN_032_B2_0	70.603125	10.52646751	A
S	Speech_019	S_TWN_001_B2_0	72.91875	10.57422448	A
S	Speech_120	S_ENS_011_XX_0	82.784375	10.79084847	A+
S	Speech_031	S_KOR_013_B2_0	75.93125	11.06940602	A
S	Speech_021	S_IDN_012_B1_2	70.0125	11.3455872	A
S	Speech_107	S_KOR_016_B1_2	72.921875	11.75786865	A
S	Speech_098	S_CHN_035_B2_0	70.7375	11.77568915	A
S	Speech_138	S_ENS_004_XX_0	79.3875	12.75046153	A
S	**Speech_038**	**S_CHN_046_B2_0**	**62.353125**	**9.651471845**	**B**
S	**Speech_061**	**S_TWN_002_B2_0**	**69.303125**	**9.836822536**	**B**



Sample			Mean Rating		Level
S/E	Sample Code	Original Code	ORS	CV	Benchmark
E	**Essay_003**	**W_CHN_009_B1_2**	**74.690625**	**9.718506101**	**A**
E	**Essay_048**	**W_PHL_064_B2_0**	**75.396875**	**10.40936679**	**A**
E	**Essay_093**	**W_SIN_005_B2_0**	**74.525**	**11.03433247**	**A**
E	**Essay_109**	**W_HKG_008_B2_0**	**73.85**	**11.98636279**	**A**
E	**Essay_030**	**W_CHN_160_B2_0**	**70.0125**	**12.58790433**	**A**
E	Essay_121	W_ENS_003_XX_1	71.903125	12.78956764	A
E	Essay_124	W_SIN_006_B2_0	76.509375	12.95256626	A
E	Essay_126	W_IDN_177_B2_0	70.39375	12.99119614	A
E	Essay_099	W_SIN_014_B2_0	74.49375	13.62962203	A
E	**Essay_031**	**W_TWN_007_B2_0**	**69.5875**	**9.67442926**	**B**
E	**Essay_070**	**W_IDN_178_B2_0**	**68.003125**	**11.31149677**	**B**

Sample quality levels were decided from the mean overall rating scores (ORS): A: >=70%, B: >=60%, C: =>50%, D: =>40%, E: =>30%, and F: <30%. Then, among all the samples belonging to each level, five samples with the smallest CV values were chosen as candidates for the benchmark sample, which appears in bold in the table.

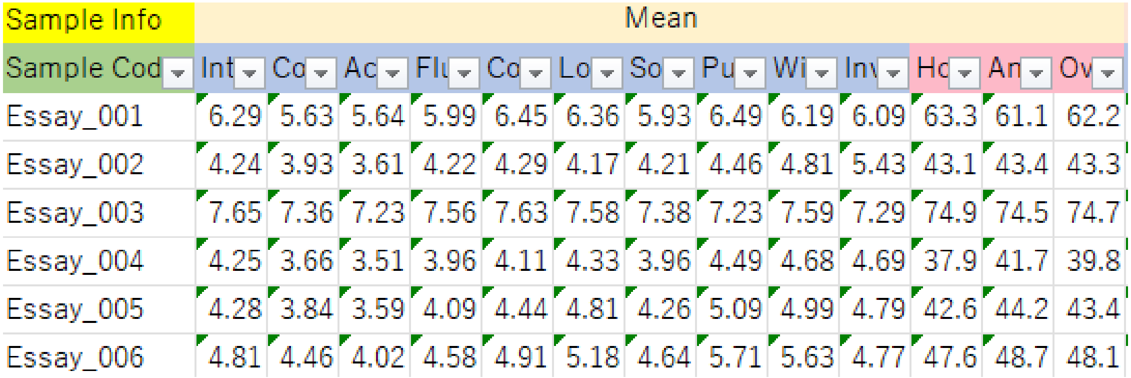
Here, we pay attention to the Level A speech samples. Among 140 samples, 13 were classified as Level A (ORS: =>70%). Then, five samples showing the lowest CV values (Speeches 040, 039, 058, 029, and 009) were chosen as candidates for the Level A benchmark. Regarding speeches, we have three samples whose ORS reach 80% (A+), but they are regarded as a part of Level A. As the number of Level F samples was seven in speeches and only one in essays, benchmarks were not chosen for Level F.

3.3.6 Check Sheet

The check sheet illustrates how each value shown on the summary sheet is calculated. In the left half, cells are filled with numbers (rating scores), while cells in the Mean, SD, and CV sections are all filled with functions. Therefore, users can see how each value is calculated.

Figure 17

Information in the Check Sheet



Sample Info	Mean											
Sample Cod	Int	Co	Ac	Flu	Co	Lo	So	Pu	Wi	Inv	Ho	An	Ov
Essay_001	6.29	5.63	5.64	5.99	6.45	6.36	5.93	6.49	6.19	6.09	63.3	61.1	62.2
Essay_002	4.24	3.93	3.61	4.22	4.29	4.17	4.21	4.46	4.81	5.43	43.1	43.4	43.3
Essay_003	7.65	7.36	7.23	7.56	7.63	7.58	7.38	7.23	7.59	7.29	74.9	74.5	74.7
Essay_004	4.25	3.66	3.51	3.96	4.11	4.33	3.96	4.49	4.68	4.69	37.9	41.7	39.8
Essay_005	4.28	3.84	3.59	4.09	4.44	4.81	4.26	5.09	4.99	4.79	42.6	44.2	43.4
Essay_006	4.81	4.46	4.02	4.58	4.91	5.18	4.64	5.71	5.63	4.77	47.6	48.7	48.1

4. Conclusion

As a unique and considerable size of L2 assessment dataset, the ICNALE GRA could be utilised for a variety of studies in applied linguistics and its related fields.

For example, researchers in testing and assessment fields could explore the rating data from 160 raters to discuss the intra- and inter-rater reliabilities and the possible effects of rater backgrounds on the rating performance. They could discuss a variety of research questions such as which of the native-speaker raters and non-native-speaker raters are more lenient, which of the English teachers and the others are more reliable raters, how non-native-speaker raters' L1 backgrounds influence their rating, and whether raters become more lenient when assessing the output of the students sharing the same L1 backgrounds, for instance.

Researchers and practitioners in TESOL could identify the students' speeches and essays that are highly rated by many raters with varied L1, regional, and occupational backgrounds. This would be a good model to present in class. Also, a comparison between highly and lowly-rated outputs would reveal concrete hints for a better speech or essay.

Those interested in the development of automated scoring engines, one of the booming research areas, could try various types of regression modelling using the overall rating scores given to samples as dependent variables and the frequencies of lexicogrammatical features obtained from them as dependent variables.

Finally, scholars in learner corpus research could sophisticate the analytical framework called a contrastive interlanguage analysis (Granger, 1996, 2015) by using high-quality learner samples as a substitute for the native-speaker outputs as a yardstick of comparison. As Gilquin (2022) emphasises, scholars of learner corpus research need to pay more attention to the diversity of the yardstick data. Learner outputs whose quality is guaranteed by a numberless group of raters with varied backgrounds could be a good candidate for a new yardstick.

Using the pre-release data of the ICNALE GRA v1.0 (120 raters in total), the author conducted several case studies. For instance, Ishikawa (2023a) discussed the intra-and inter-rater reliability, interrelations between rating categories, the effect of rater background variables (pp. 177–183), the possibilities of automated assessment (pp. 184–190), the quality of L1 English native speaker outputs, the possibilities of using top-level learner samples as a new yardstick, and the possible use of top-level learner outputs in classrooms (pp. 190–201). Ishikawa (2023b) discussed the possibility of using the top-level learner outputs as an addition to or a substitute for the traditional native-speaker outputs as a reference in contrastive interlanguage analysis. Also, Ishikawa (2023c) quantitatively investigated the collected rating data and questioned the widely believed superiority of a native-speaker English teacher as a rater. These studies might offer a hint for the fuller utilisation of the ICNALE GRA.

Acknowledgement

This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Number 20H01282. The project began in April 2020 and ended in March 2024.

Bibliography

Gilquin, G. (2022). One norm to rule them all? Corpus-derived norms in learner corpus research and foreign language teaching. *Language Teaching*, *55*(1), 87–99.

Granger, S. (1996). From CA to CIA and back: An integrated contrastive approach to computerised bilingual and learner corpora. In K. Aijmer, B. Altenberg, & M. Johansson (Eds.), *Languages in contrast: Papers from a symposium on text-based cross-linguistic studies* (pp.37–51). Lund University Press.

Granger, S. (2015). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research*, *1*(1), 7–24.

Ishikawa, S. (2020). Aim of the ICNALE GRA project: Global collaboration to collect ratings of Asian learners' L2 English essays and speeches from an ELF perspective. *Learner Corpus Studies in Asia and the World*, *5*, 121–144.

Ishikawa, S. (2023a). *The ICNALE guide: An introduction to a learner corpus study on Asian learners' L2 English*. Routledge.

Ishikawa, S. (2023b). A new yardstick of comparison for contrastive interlanguage analysis: A study on the ICNALE Global Rating Archives. In U. Widiati, M. Hidayati, N. Suryati, Suharyadi, A. Nunuk Wulyani, I. L. Damayanti, N. A. Drajati, S. Karmina, E. L. Zen, L. Hakim, & Prihantoro (Eds.), *Advances in social science, education and humanities research: Proceedings of the 20th AsiaTEFL-68th TEFLIN-5th iNELTAL conference (ASIATEFL 2022)* (pp. 607–619). Atlantis Press.

Ishikawa, S. (2023c). Effects of raters' L1s, assessment experiences, and teaching experiences on their assessment of L2 English speeches: A study based on the ICNALE Global Rating Archives. *LEARN Journal: Language Education and Acquisition Research Network*, *16*(2), 411–428.

Appendix

Rating Guide for The ICNALE Global Rating Archive (ICNALE GRA)

After reading this guide, you need to take a check test. If the test score is eight or lower than that, you are required to take a test again. Please read this carefully before taking the test.

1. What is the ICNALE?

The ICNALE is a collection of essays and speeches by college students in ten countries/ regions in Asia and English native speakers. The ICNALE is the largest corpus of Asian L2 English learners.

2. What is the ICNALE GRA?

The ICNALE GRA is a collection of the evaluation data of spoken and written samples included in the ICNALE. The same samples will be evaluated by approx. 100 raters having different L1 and occupational backgrounds.

3. The participants of the ICNALE Project

More than 4,000 students and native speakers participated in the ICNALE project. Among them, 140 participants were randomly chosen for the ICNALE GRA project. Participants' backgrounds are kept secret from the raters.

4. Two kinds of tasks

There are two kinds of tasks.

Speech (Roleplay)	Essay
A participant was required to play the part of a college student wishing to continue their part-time job. They needed to persuade their supervisor, who firmly believed that college students should never have any part-time jobs. NB: Raters will listen to an initial 90 seconds of a roleplay. Raters do not need to evaluate the output of an interviewer	A participant was required to write 200-300-word essays about the topic "*It is important for college students to have a part-time job.*" They needed to show examples and details to support their claims. NB: Raters will read the whole essay and evaluate the quality.

playing the role of a college supervisor.	

5. Two kinds of ratings

First, raters will conduct an overall (holistic) evaluation, and after that, they will conduct an analytical evaluation based on ten evaluation criteria.

6. Rating Scale

The 0-100-point scale is adopted for overall evaluation, and the 0-10-point scale is adopted for analytical (criterion-based) evaluation.

		Overall Evaluation (0-100)	Analytical Evaluation (0-10)
Positive ↑	Awesome+	100	10
	Awesome	90(-99)	9
	Excellent+	80(-89)	8
	Excellent	70(-79)	7
	Good	60(-69)	6
Neutral	Average	50(-59)	5
Negative ↓	Fair+	40(-49)	4
	Fair	30(-39)	3
	Poor+	20(-29)	2
	Poor	10(-19)	1
	Unacceptable	0(-9)	0

[Middle point]

The average, namely, the middle point on the scale, is 50 (or 5). The value higher than 50 (5) should be positive, while the value lower than 50 should be negative.

[Referential standard]

Each sample should be evaluated from the viewpoint of English as a Lingua Franca (ELF), a type of English used mainly for professional communication between non-native speakers having different mother tongues (e.g., between L1 Japanese and L1 Thai speakers). According to recent research, more than 75% of English communication in the fields of business and research occurs between non-native speakers. Raters are expected to fully understand the status of English in the current world.

Therefore, "Excellent" in the rating scale, for example, should be understood NOT as being excellently close to English native speakers but as being excellent as a professional ELF speaker.

[Variance in rating]
It is important for raters to rate each speech/essay with sufficient range and variance. Please note that your scores need to vary well from "0 to 100" or "0 to 10". If you give similar scores to most of the speeches/essays, you will be requested to redo your rating.

7. Overall Evaluation
To what extent do you think the sample is close to an ideal ELF speech/essay? Raters have to examine each sample and decide the score (0-100) based on the overall judgment of its quality as a professional ELF output. Please note that "90 points," for example, should be given to someone who you think is a 90% ideal professional ELF user, not to someone who you think is 90% close to English native speakers. Also, please note that the middle point is 50.

8. Analytic Evaluation
Then, raters have to examine each sample based on the ten criteria shown below and decide the score (0-10) for each of them.

The criteria are classified into language-related criteria, content-related criteria, and attitude-related criteria, though they are often overwrapping.

(Mainly) Language-related	(Mainly) Content-related	(Mainly) Attitude-related
(1) Intelligibility	(5) Comprehensibility	(9) Willingness
(2) Complexity	(6) Logicality	(10) Involvement
(3) Accuracy	(7) Sophistication	
(4) Fluency	(8) Purposefulness	

(1) Intelligibility
To which extent can you "decode," namely, verbally understand what is said/written? In speech evaluation, factors such as pronunciation and intonation will influence the degree of intelligibility. In essay evaluation, factors such as spelling and sentence structure may influence it. Please note that intelligibility, which concerns the understandability of the language, should be discriminated from comprehensibility, which concerns the understandability of the content. You may sometimes find a speech/ essay that is intelligible but not comprehensible, such as a logically nonsense statement. Meanwhile,

you may usually not find a speech/ essay that is comprehensible but not intelligible because if the text cannot be decoded, its content cannot be conveyed.

(2) Complexity

To what extent do you think the speaker/ writer uses morphologically and/or semantically complex words, phrases, expressions, constructions, and grammar? Complexity is seen at many levels of language. For example, “I speculate...” usually sounds more complex than “I think” (Vocabulary). “It is speculated that...” may sound more complex than “I speculate” (Voice, Construction). “If I were a bird” may sound more complex than “If I am a bird” (Subjunctive, Grammar).

(3) Accuracy

To what extent do you think the sample is error-free in terms of vocabulary and grammar? In addition, you should examine the elements such as pronunciation and intonation in speech evaluation and those such as punctuation in essay evaluation. Please note that you should ignore minor and only-once errors, which may be mistakes rather than errors. Please note that the standard for evaluation should be a proficient non-native ELF speaker, not an English native speaker.

(4) Fluency

To what extent do you think the speaker/writer is fluent in the speeches/ essays? Fluency needs to be evaluated in two ways: (a) fluency and (b) disfluency. If someone talks/writes more, the fluency score should increase, while if s/he uses more disfluency markers, the score may decrease. Disfluency markers include fillers (uh, well, oh, hmm), pauses, false starts (I thin... thin... no, I thought...), etc. in speeches, and unnecessary connectors (and, but, so, because) and semantically empty phrases (such as “I think” most typically), etc. in essays. Please note that using these disfluency markers once or twice usually does not cause any problems in communication.

(5) Comprehensibility

To what extent can you understand the content of the speech/essay? Please note that comprehensibility, which concerns the understandability of the content, should be discriminated from intelligibility, which concerns the understandability of the language. If a speaker/writer presents a logically reasonable idea, the score should increase.

(6) Logicality

To what extent do you think the idea presented in the speech/essay is logical and reasonable? In speech evaluation, you need to examine whether the reasons presented by the speaker really explain why s/he needs to continue working. In essay evaluation, you need to examine whether the reasons and the conclusions are logically connected.

(7) Sophistication

To what extent do you think the ideas presented in the speech/essay are well-sophisticated,　critically thought, unique, original, and innovative?

(8) Purposefulness

To what extent do you think the speaker/writer consistently and consciously pays attention to the purpose of the task? The participant was requested to persuade a supervisor to allow them to continue working on a speech task and to show their own opinion about part-time jobs for college students in an essay. You have to examine whether the participant fully understands the purpose of the task and consistently sticks to it. Purposefulness is closely related to task completion.

(9) Willingness

To what extent do you think the speaker/writer is willing to communicate? It is possible that a participant with a limited L2 proficiency shows a high level of willingness to communicate (WTC), and it is also possible that a participant with a high L2 proficiency shows quite a low level of WTC. In a speech, factors such as the quantity of talk, the number of turn-takings, change of tones, and the use of body language may reflect the participant's WTC. In an essay, factors such as the quantity of writing, the number of ideas s/he presents, and the use of various amplifiers (e.g., "very," "surely," "definitely," "I strongly believe," etc.) may represent the participant's WTC.

(10) Involvement/Engagement

To what extent do you think the participant tries to make the hearer/reader involved in his/her discourse rather than speaking/writing one-sidedly? The factors such as the use of the second-person pronouns (e.g., "You know," "as you see," "as you expect," etc.) and mentioning the hearer/reader are usually related to the degree of involvement.

9. Short Comments

A rater needs to give short (Approximately 10-20 words) comments about the strong and weak points of each speech/ essay.

Example comments (For speeches)

Strong Points	Weak Points
Good pronunciation and good grammar control. Almost no vocabulary problems. (12 words)	Maybe need to take more initiative to complete the task. Could not fully understand the logic of the story. (19 words)
The speaker tried to persuade the listener throughout the talk. He could come up with many reasons. Good start, considering the relationship with the listener. Although she could not speak so much, she knew what she had to do.	Major grammatical mistakes were observed. Used only basic words and simple sentences.
He could answer closed questions with confidence, even though he answered only with Yes/No.	Her speech volume is small, and she pauses a lot. The given reasons were not persuasive. Should take more initiative to persuade the listener.
The speaker continuously paid attention to what she had to do. She took the relationship with the listener into consideration while talking.	Insufficient speech volume. Wrong/unnatural word usage. The given reasons were not persuasive. Can't give details. He is very passive.

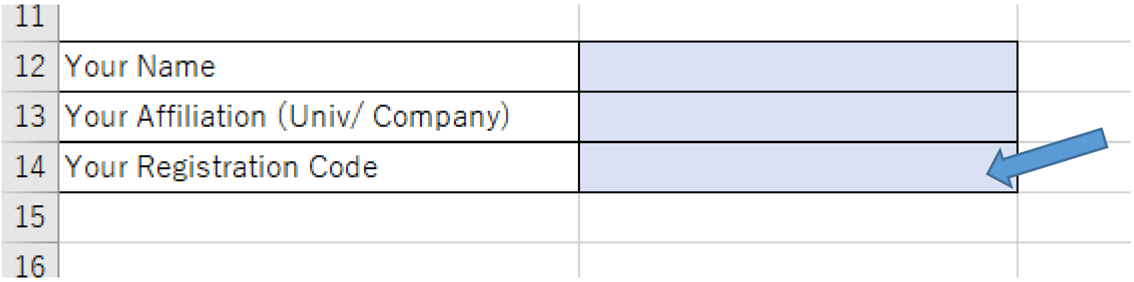
10. Obtaining the learner's speeches/essays

(Speech) Links to the sound files are shown on the Excel Sheet.
(Essay) ---Deleted---

11. Procedure

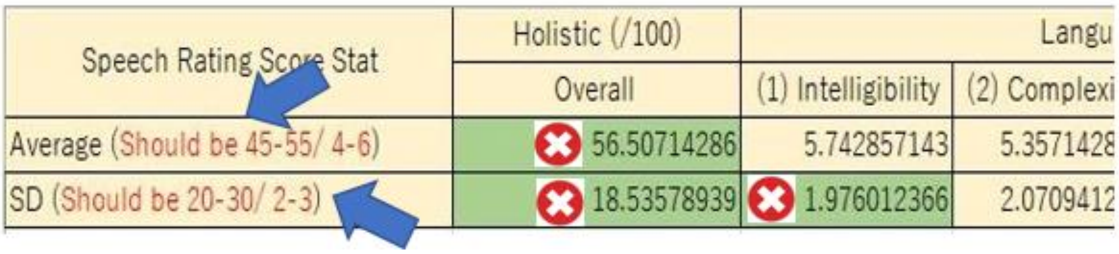
(1) First, carefully read the rating guide.
(2) Then, register your background info at the site below and take a check test.
---Deleted---
(3) If your score is eight or below, check the guide again, and please take a retest soon.

(4) Please enter the registration code, which appears on the screen after your submission, to the Rating Sheet (an Excel file) and begin the rating.



11		
12	Your Name	
13	Your Affiliation (Univ/ Company)	
14	Your Registration Code	
15		
16		

(5) When you finish rating all of the 140 speeches or essays, please confirm that both the average and standard deviation (SD) are within the preset range, which are shown on the first sheet of your Excel File. If they are not within the range, please correct your rating scores before you send them.



Speech Rating Score Stat	Holistic (/100)	Langu	
	Overall	(1) Intelligibility	(2) Complexi
Average (Should be 45-55/ 4-6)	✖ 56.50714286	5.742857143	5.3571428
SD (Should be 20-30/ 2-3)	✖ 18.53578939	✖ 1.976012366	2.0709412

In this case (see the fig. above), the average of the overall score, the SD of the overall score, and the SD of the intelligibility score are all out of the range. These should be fixed.

(6) Then, send your Rating Sheet to the project leader at your institution.

(7) Please note that if we find something wrong or inappropriate with your rating, we may ask you to redo it.

(8) If you have a question, do not hesitate to ask Dr. Ishikawa.